Review Team #3

September 8, 2006

Review of 'Developing methods to reduce bird mortality in the Altamont Pass Wind Resource Area' by K. S. Smallwood and C. Thelander (2004)

This final report is a revision of an initial report that considers the work of Smallwood and Thelander (2004). I have responded to their comments regarding broader issues raised by all of the reviewers first. I then consider their responses to my specific comments in my initial report. Changes in my initial report are represented in italics. In some places, I state the authors' response and then further explain my point. In other cases, my initial statement and their response require no further comment. Clearly, the authors considered the reviewers' comments and thoughtfully explained how they would modify their methods if future work was to be performed.

Pseudoreplication was the first broad issue raised by reviewers to which the authors responded. However, the first 2 paragraphs under this section (bottom of page 5 and top of page 6) are not about pseudoreplication, rather they refer to nonrandom selection of turbine strings. Hence, I will address these paragraphs and related material first and then consider the pseudoreplication issue offered in the third and fourth paragraphs under the first subtitle (i.e., second and third paragraphs of page 6).

Nonrandom sampling of turbine strings

Smallwood and Thelander agree that random sampling initially would have been preferable, but they disagree with the premise that statistical inference depends on samples being randomly selected. Thompson (*Sampling* 2002:2) states 'Sampling is usually distinguished from observational studies, in which one has little or no control over how the observations on the population were obtained.' Subsequently, he emphasizes that random sampling can avoid many of the factors that make observational data "unrepresentative". The take home message here is that Smallwood and Thelander have performed an observational study. Observational studies are commonly performed in the biological sciences. However, as such, they do have limitations. One of these limitations is the degree to which defensible inferences can be made. In my initial review, I questioned the inferential capability of this work because of the nonrandom sample of turbine strings. I encouraged the authors not to infer to the larger APWRA because they do not have a random sample from the APWRA. While this is not their fault (they surveyed all turbines to which they were granted access), nevertheless, its implications remain. A nonrandom sample does not necessitate that their results are not representative. For instance, Cochran (*Sampling Techniques* 1977:10) states that under the right conditions, some common types of nonrandom sampling can give useful results. However, he further states that the only way of examining how good one of them may be is to compare the results to a situation in which the results are known, either for the whole population or for a probability sample. If one acknowledges that inference should only be made to those turbines which were surveyed, the question then becomes does the time period in which they surveyed reasonably represent the temporal nature of the system, for which I have no answer. However, at least they have multiple years of surveys and multiple surveys within each year for the portion of turbines surveyed.

On page 7, Smallwood and Thelander state they could not have purposely biased selection of turbines- I have no such suspicions. I am only concerned about unknown sources of bias due to nonrandom selection. The authors selected turbines systematically to intersperse searched and unsearched turbines. This is a common approach used in biological studies. However, a systematic approach can be a random method if the initial starting point is randomly selected from within the collection of units to be sampled, thereby allowing all intervals of sampling units to have a nonzero chance of selection. A systematic random sample has the same structure as a cluster sample, and can be treated as a cluster sample of size one. Variance estimation is not straightforward for such samples, unless the systematic sample is replicated (i.e., more than one collection of spaced units is selected). I encourage the authors to read chapter 12 in Thompson's book (Thompson 2002).

Pseudoreplication

To properly discuss this issue, one must identify the sampling or experimental unit. First, an experiment has not been performed. Treatments have not purposely been applied to subjects with the notion of eliciting a response. Second, has a sample of turbines strings been examined? The answer is yes, but it is a nonrandom sample of the entire APWRA (to be discussed subsequently). What was measured? On page 47, the authors indicate that turbine strings were surveyed using transects. In this sense, the turbine strings are the sampling unit. Turbines within strings might be considered subsamples, depending on the situation. Certainly, individual turbines within strings are correlated spatially and temporally in terms of measurements made. Thus, use of individual turbines must be considered carefully so as not to inflate sample sizes in statistical tests.

Extrapolation

Extrapolation of the mortality estimates may be looked upon as acceptable by some researcher, but statisticians as a whole do not encourage such practices because extrapolation is inherently problematic. The authors state they believe the extrapolation was reasonable, and attempt to justify it based on the interspersion of turbines searched, etc. 'Trust me' is not the basis of good science. Point estimates are of little value in estimation. Suppose I was to ask the authors to develop a 90% confidence interval around their extrapolation, could they do so? In regression modeling, there exists a defensible means of doing so, but only within the range of the data collected, so why then are the authors so willing to go beyond the range of their data in this case?

The authors clarify that they did not extrapolate the results of fatality associations to wind turbines outside the measured set. My question is then, to what population are the statistical tests referring? I suspect the answer might be that their population is the collection of mortality strikes annually at the wind turbine set. If so, the important question then becomes were the turbines searched enough throughout the year to adequately represent it in its entirety, and if not, were the times within the year randomly selected so as to infer to an annual timeframe.

Confounding

Their response to this problem was basically an acknowledgment that it may have occurred. They give an example of how they could reduce some of these effects if given the opportunity to revise the report. While I agree that not all instances of confounding might have caused misinformation, I remain concerned that some of the conclusions drawn might be inappropriate because of lack of attention to the issue. My recommendation to the authors is to identify the degree to which certain variables are confounded, i.e., look at the 'treatment structure' of the turbine strings and identify situations in which confounding can be reduced or eliminated. When confounding cannot be averted, clearly state this whenever a test is performed and when interpreting the test result. For instance, they could say, 'it remains unclear whether variable B or C is actually associated with this high mortality rate'.

Multiple Comparisons, Multivariate tests, and use of Chi Square Tests

I agree with the authors that alternative analysis methods will not eliminate all of the problems associated with the data set, e.g., nonrandom sampling, lack of replication, etc. The selection of variables was under their control as were the analytical methods and it is here that more thought should be placed. I recommend the authors consider use of some form of multiple regression in combination with model selection methods. The latter may rely on information theoretic approaches, Mallow's Cp, likelihood ratio testing, etc. Specific methods should be discussed in consultation with a statistician. For instance, logisitic regression could be used with the response being two categories (deaths occurred or not) vs. several explanatory variables, including effort. Polytomous logistic regression would allow for more than 2 categories, so they could have high mortality, low mortality and none as the response variable. An alternative would be to use generalized linear models such as Poisson regression, in which effort could be an explanatory variable, or using multiple regression with a response variable of mortalities per unit effort. I understand the idea behind the authors' use of the Chi square goodness of fit testing, but believe this approach is inferior because it is only considering one variable at a time. The authors later sum over individual variables, but there is no basis for doing so under the sampling design and replication with which they had to work.

Type I error

The authors response to this criticism has identified a misunderstanding of what Type I error and P-values represent. First, probability is prospective. Second, whether or not a Type I error has been committed does not depend on the P-value (other than the test had to be rejected, or below the stated alpha). If I were to perform 100 statistical tests using an alpha level of 0.05 in each, then it would be very reasonable to have 5 rejections even when none of the null hypotheses were false. The authors' response is that many of their P-values were small. This does not change the above interpretation; notice I did not even mention the P-values for the 5 rejections. A P-value is not a measure of effect size and it is not legitimate to adjust an alpha level based on a p-value. Another viewpoint from the authors is that they are willing to live with some type I errors. For example, suppose 100 tests were performed and 15 were rejected. Again, let's assume 5 were rejections of true nulls. The important question would then be which ones are false rejections and which are not. The authors are less concerned about answering this question than I expected.

What follows is an edited version of the initial report

General Summary

Smallwood and Thelander do an excellent job of describing background information and previous literature regarding bird mortalities at wind farms. A strong case is made for the relevance of the work they performed. Their objectives were multi-faceted and involved examining bird behaviors, raptor prey availability, turbine characteristics, landscape features, and bird mortalities. Each of these components is a substantial endeavor and the authors are to be commended for examining this multitude of factors. Clearly, the authors used a 'heads up' approach that involved making observations, and then attempting to collect data and evaluate hypotheses formulated from these observations. For example, they expanded their research to examine fossorial animal burrows and effects of rodent control on bird mortality.

While I believe they have substantial information regarding all of their objectives, the details of how to extract this information and to what the information refers to become the most important considerations I have regarding the report. I believe that alternate analytical methods should be explored for examining their data, which may lead to different interpretations of their data. Even if the interpretations remain similar, I would have greater confidence if their analyses were improved upon. Furthermore, I would suggest extensive consultation with a statistician to assist in such an endeavor. *Smallwood and Thelander have indicated a willingness to do so if allowed to revise the report.* It is my belief that a great deal of thinking and interaction with a statistician is needed to ensure that the information contained in this impressive data set is properly interpreted. Given the variety of data collected, including different scale issues, multiple response and explanatory variables (they refer to as dependent and independent, respectively), various data types (categorical, continuous, ordinal), etc., I believe many hours of interaction is needed to identify the most prudent analysis approach for each of the chapters. I would anticipate spending well over 100 hours on such a project in a statistical consulting role.

My criticisms that follow are meant to be constructive, not destructive, and I hope that the authors will take them as such. I likely have misunderstood their intentions at various places in their report and I assume the authors will correct my thinking on these portions of my assessment. Thus, I suspect some of my criticisms can easily be rectified. There was an opportunity to submit questions and have them answered by the authors. However, in my experience, the importance of the nonstatistical aspects of statistical consulting cannot be understated. For example, direct verbal interaction is the best means for arriving at a consensus understanding of the material. Use of reflective listening is a very useful technique for making sure everyone is 'on the same page'. The format for this review did not allow such interaction; however, I recognize the utility of the process employed in this review.

Strength of inference is determined by the study design. Strength of evidence is determined by the data alone. In this case, the study design does not lend itself to a strong inferential setting for two reasons. First, the variables of interest are likely confounded because turbine string placement was not designed with their study objectives in mind. Hence, there are many factors that potentially affect the response variables of interest that are not separately estimable because all combinations of explanatory variables are not represented on the landscape. For example, if one was only interested in the effects of aspect (north, east, south and west) and tower height (say 4 categories), then one would need 16 tower-aspect combinations represented, with replication, for sufficient estimability of main and interactive effects. Thus, this study lacks replication of the set of 'treatment' combinations of explanatory variables being examined (see Johnson 2002 for a discussion on the importance of replication in wildlife research).

Smallwood and Thelander indicate some misunderstanding of this issue in their response, in which they state the above problem is the reason they tested one variable at a time. There may be 'replication' of most of the variables, but there is substantial confounding of these variables. When variables are confounded, their effects are not separately estimable, therefore one-variable-at-a-time analyses can be misleading. Tests can be conducted one variable at a time, but such tests do not escape the problem of confounded effects. For example, if explanatory variable A affects the response variable C, but explanatory variable B does not, a test of B's effect on C is likely to be detected if A and B are confounded, even though it has no effect, because it is confounded with A, which does have an effect. Ultimately, their approach may have led to several spurious effects, upon which management recommendations might be made. Confounding is not something that can easily be addressed, it must be seen as a problem that limits inference. In addition, there are good reasons why multiple regression is commonly used (two or more explanatory variables in one model) as opposed to multiple versions of simple linear regression using one explanatory variable at a time. Multiple regression enables one to assess the impact of one explanatory variable in the presence of other explanatory variables. More complex models typically are more helpful in providing sufficiently precise predictions. When using multiple regression, one must be mindful of potential multicolinearity problems when two or more explanatory variables are highly correlated. Multiple regression is not an example of multivariate methods, the latter refers to the situation in which there are multiple response variables, not multiple explanatory variables. Their approach of univariate Chi-square goodness of fit tests to identify individual variables of interest, and the associated metric of the ratio of observed to expected as a measure of effect (which they later sum over for certain variables), is more of an ad hoc approach that does not readily draw upon formal statistical procedures for model selection and testing. As stated earlier, I believe a substantial amount of communication should take place with a statistical consultant to best arrive at an understanding of some important statistical issues, which I can only assume are poorly understood at this point.

Second, because the sample of sites studied were not randomly sampled any inferences based on statistical tests are not statistically defensible. The collection of sites studied represents those locations to which the researchers were allowed access or which were accessible due to environmental conditions. I do not fault the researchers for this shortcoming; they surveyed what was made available to them. However, it does weaken the strength of inference they are able to impart and I suggest they reconsider the context of many of their statements of inference. For instance, hundreds of tests (Chi-square, analysis of variance, tests of correlation, etc.) are performed in the report, but the population(s) to which inferences are made is at best, unclear, and at worst, inappropriate.

In response to this criticism, the authors clearly state that they had no investigator bias. Randomization is important to avoid any source of bias, many of which are unknown to the investigator (I never meant to imply there was an investigator bias). For instance, the collection of turbines surveyed may underrepresent or overrepresent a particular landscape attribute, which may affect likelihood of strikes. In my initial report, I acknowledged that the nonrandom sampling was partly unavoidable and does not necessitate that their sample of turbine strings is poor. They surveyed all of the turbines in the first set of turbines that were made accessible. However, in the second set, they selected a portion systematically. Systematic sampling is commonly in biological studies because it forces a spatial evenness to the sampling. Before sampling has begun, one can envision a set of sampling units, called a primary unit, separated by a specific interval. As long as the initial selection of the primary unit is randomized, a random sample has been achieved. Often, researchers use their judgment to select the primary unit, which needlessly results in a nonrandom sample. It is unclear how the authors selected the turbine strings in the second set. In any case, if the population being inferred to is the entire APWRA, a nonrandom sample of this area has been taken. One approach the authors could take would be to simply describe the results from their sample. There would be no uncertainty in these descriptive statistics, although there are detectability issues that would imply the estimates presented are below the true mortality number. Such an approach would eliminate the extrapolation of mortality estimates to the entire APWRA that I criticize later. If the authors choose to infer to the entire APWRA, then such estimates cannot be defended statistically, i.e., acceptance depends on the judgment of their audience.

In some instances, the authors did have the opportunity to randomly sample from the collection of turbines to which they had access. Granted, this target population is not the entire APWRA, but random sampling would have justified inferences to a larger population. For example, when studying rodent burrow locations, the authors have not indicated that a random sample of turbines or turbine strings was selected from those available; they stated turbine strings were 'arbitrarily selected'. Thus, I am not clear as to what population the inferences are being made.

Hypothesis testing is one form of statistical inference. Estimation of parameters is another form of inference the authors employed. Limitations due to nonrandom sampling must also be considered in this context. For example, in estimating bird and raptor mortality (e.g., see executive summary) for the Altamont Pass Wind Resource Area (APWRA), a nonrandom sample of turbine locations was surveyed from the entire APWRA. While I do not blame the researchers for this circumstance, it does remove the defensibility of statistical inferences. Random sampling does not necessarily ensure a representative sample, it only ensures that on average, in repeated sampling, the population will be well-represented (a conceptual property). That the authors have surveyed a majority of turbines provides some support for the notion they have a good representation of the population, however, only approximately 28% of the turbines were measured at least 3 years. If the authors can make the case that their sites are representative of the larger APWRA, then perhaps the scientific inference (not statistical) being made will be acceptable to all. I realize that many of the environmental attributes might be unknown for those sites not visited, but I would think that the turbine/tower features would be known for all turbines in the APWRA and could be considered within the context of those that were sampled. In their executive summary, they estimate the number of raptors (and all birds combined) killed annually in the APWRA. I suggest they *(changed the wording here from Personally, I would*) refrain from making such inferences and limit estimates to the area specifically surveyed, given the strength-of-inference limitations due to nonrandom sampling and the uncertainty in the 'adjustments' they make in their estimation process. Their projections for all wind-generating facilities in the United Stated (page 86) should also be considered as extrapolations without much credence.

They further state that the risk to birds has increased substantially over the past 15 years, indicating a formal trends analysis. This support for this statement is not satisfactory unless more information is given on consistency of detection rates over this time period. Note there are many factors that can cause inconsistencies in observed counts over time, including surveyor differences, environmental differences, and animal behavior differences. The authors did suggest that birds may have altered their behavior in response to the presence of turbines in the area. I suspect that there are many survey methodological differences over this time span as well. *The authors indicated a willingness to replace chapter 4 with a more rigorous treatment of the issue if given the opportunity.*

There is also a temporal component of their sampling which is clear in the general sense that turbines were surveyed a different number of occasions, etc., but the details of how often/when specific turbines were surveyed and how this might affect interpretations are not explicit. For instance, if tubular towers were surveyed more often and at times of higher bird abundance than other tower types, then greater mortalities observed may merely be a function not of the tower type, but greater effort and timing of greater bird abundance. Some sites (1526 turbines) were surveyed over a 3-year period, with a staggered entry of turbines over that period, so turbines had differing numbers of surveys executed within that period. For example, on the middle of page 48 they state they were limited to 685 turbines through 1999. Another set of turbines (2548) were surveyed over a 6-month period. If I understand their methodology, the authors computed estimates of availability that account for survey effort by placing landscape and turbine features on a relative (proportional) basis. However, in most use-versus-availability studies I have seen, 'availability' is assumed known, when it is almost always estimated. What is 'available' from the human perspective of what is on the map, is not necessarily available to the animals of interest, even if they are highly mobile, because they may not have such a map in mind when making decisions. Alldredge et al. (1998) provide a nice overview of statistical approaches to resource selection studies that nicely clarifies the set of assumptions underlying such analyses. If the authors do not modify their analytical methods, which I strongly recommend, then at the least they could more explicitly state the assumptions underlying their approach, the likelihood that the assumptions are valid, and the ramifications if not valid. *The authors have indicated a willingness to do so if given the opportunity to revise the report.*

Another important design component includes consideration of what the multiple competing hypotheses are and how best to discriminate among them. When possible, readers of this report should be informed of the hypotheses under consideration and how the sampling scheme used can discriminate among these hypotheses. For example, there is an entire section of work in chapter 2 that examines the distances of 'small' versus ‘large’ birds from the turbine, yet there is no explanation of why the data are being partitioned as such, i.e., what the hypothesis is, and how this partitioning relates to assessing the efficiency of their search radius. Another example of the importance of considering one’s hypotheses is demonstrated in chapter 6. The objectives are clearly stated, but the *a priori* hypotheses regarding the effects of rodent control are not stated. They allude to the notion of ineffectiveness, but I would like to see explicit hypothesis statements. Lacking the benefit of observations, my scientific hypothesis would be that increased intensity of rodent control results in fewer raptors, thus lowering susceptibility to strikes and thus lowering observed fatalities. *The authors stated that they would strive to clarify their hypotheses if given the opportunity for revision.*

How best to discriminate among hypotheses is an important design consideration. For example, in studying the effects of rodent control, the authors did select sites with a wide range of observed raptor mortality and rodent control intensity. This approach enhanced the ability to distinguish among competing hypotheses. However, they did not random sample turbine strings according to these features (e.g., a stratified design), thus limiting the defensibility of inferences made.

Care must be taken when attempting to demonstrate ‘treatment’ effects. For instance, the treatments must be effective in their application. My understanding is that rodent control was aimed specifically at eliminating ground squirrels, but not other species (e.g. pocket gophers). Clearly, if one species is targeted, that does not necessarily imply a significant reduction in overall prey availability, in fact, it may increase it. *On the other hand, elimination of one species might focus raptor presence in other locally defined areas with other prey species. The important question would then be are these areas more or less susceptible to bird strikes?* Thus, I question if the rodent control ‘treatments’ were substantial enough to observe an effect. *In their response, the authors assured me that the treatments were effective in killing ground squirrels*. Rodent control might be effective in reducing raptor susceptibility if prey availability (in total, not just ground squirrels) is substantially reduced and this reduction is discerned by raptors. How best to isolate treatment effects is also an important component in identifying treatment effects. In applying the rodent control treatments, were other variables that influence raptor mortality controlled? If not, not, then confounding effects may make it difficult to isolate the effect of the rodent control treatment. In summary, I suggest that the efficacy of the management actions taken be considered before discarding their usefulness. Similar arguments apply to topics such as benefits of perch guards, etc. Smallwood and Thelander have made decisions and recommendations based on observations from considerable survey effort. Again, I believe there is tremendous value in their data, interpretations of which must be carefully considered for that value to be realized. *In responding to this statement, the authors state all of their conclusions were based on evidence. While I agree that they are using evidence (data), the details of corrected interpretation of that data become important in determining validity. For instance, the presence of confounding effects may indicate a variable has an effect when in fact it does not, or vice versa (one variable may counteract another). I was simply stating that a rigorous experiment should be performed that allows isolate of specific treatments to identify treatment effects. I do not believe that such an experiment has been correctly executed as of yet.*

Regarding strength of evidence, the authors have completed a notable amount of work. Having surveyed the majority of turbine locations, some of which were surveyed multiple years, I believe there is substantial information to be discerned from the collected data. The key to harvesting that information is proper context and hard thinking about what metrics make sense, appropriate use of statistical tests, placing outcomes of statistical tests in the context of biological relevance, etc. In attempting to determine causal factors of bird mortalities, the authors surveyed locations around turbines in the APWRA that were accessible. They identified all known bird strike mortalities (approximately 1045 if the number due to unknown causes is removed) and examined associations of various turbine/tower, landscape, and environmental factors with these mortality counts. Thus, a retrospective observational study has been performed with the intent of determining causal relationships. I believe most statisticians would agree that establishing causation requires a more rigorous approach in study design. Romesburg (1981) stated that causation requires more than correlative evidence; one must eliminate other possible causes and must demonstrate similar associations that are plausible over a wide range of circumstances. Many conclusions stated in the report are plausible, but I am not convinced that causation has been established anywhere in the report.

Their analysis approach was generally that of examining associations one variable at a time. Thus, hundreds of computations resulting in hundreds of statistical tests were performed. One of the problems with such an approach is that there are likely to be spurious effects. The probability of falsely rejecting a true hypothesis at least once is essentially one when more than 50 tests are performed. Another problem with examining one variable at a time is that there are likely to be confounding effects unless all combinations of factors (34 variables) are represented. For example, if certain turbine models tend to be at higher elevations, or appear on canyons more often, then greater than expected raptor mortalities may be due elevation or landscape considerations rather than turbine model itself. Alternatively, it may be that an interactive effect exists, such that the combination of various features increases the potential for strikes. The one-variable-at-a-time approach to analysis can mask such outcomes. The authors recognize the potential for confounding effects in other studies (see appendix B, page 2), and even admit a small portion of their study (rodent control in chapter 6) is prone to confounding, yet they fail to see the possibility of confounding effects in the majority of their one-variable-at-a-time analyses. I am confident that there are several reasonable approaches that can be made to analyze their data considering multiple explanatory variables simultaneously. For example, I suggest that logistic regression or Poisson regression be considered for some types of analyses. These modeling approaches will allow for multiple explanatory variables and will not necessitate the data reduction that has been used to categorize certain continuous explanatory variables.

In summary, I believe there are specific design flaws which limit the validity of inferences to a larger population (which in some cases is not clearly defined). Reliance on descriptive information may not be seen as ‘scientific’ but I disagree with the notion one must make inferences to have useful information. Second, I suggest the authors rethink their analytical approaches. More appropriate methods of analysis would strengthen my belief in their stated outcomes and recommendations. Extensive consultation with a statistician is recommended.

(I deleted a section here that began as 'If I were considering this for publication…' because the point was moot.)

References:

Alldredge, J.R., D.L. Thomas, and L.L. McDonald. Survey and comparison of methods for study of resource selection. Journal of Agricultural, Biological, and Environmental Statistics. 3:237-253.

Johnson, D.H. 2002. The importance of replication in wildlife research. Journal of Wildlife Management 66:919-932.

Romesburg, H.C. 1981. Wildlife science: gaining reliable knowledge. Journal of Wildlife Management 45: 293-313.

Executive Summary

Specific Comments

In describing their approach, the authors state they presented mortality estimates as ranges, where the lower end was adjusted for likely outside of their search area, and the upper end was adjusted for fatalities missed due to undetected carcass removal. I would consider both of these to be upper-end adjustments, actually using both simultaneously would provide a higher upper end. The lower end would be represented by unadjusted values *(Smallwood and Thelander acknowledge this in their response)*. The upper-end estimates must be interpreted with caution. The adjustments made are based on detectability estimates from other studies that are likely to be study-specific for a variety of reasons. In addition, the range for the estimated number of raptors (and all bird combined) killed annually is given for the entire APWRA. I suggest the authors consider the accuracy of their estimator given the design weaknesses at the larger scale of sampling the APWRA and at the smaller scale of bird carcasses detected. They further state that the risk to birds has increased substantially over the past 15 years, however, I am not convinced that the data have been collected in a manner that allows for trend estimation due to likely inconsistencies in survey methods, personnel, effort, animal and environmental differences.

Justification for their defined metric of mortality as mortalities per megawatt (MW) per year *was not convincing (I modified this ending from 'is not properly stated'). As with any metric, the variable of interest must be clearly defined.* They give the reason of 'to avoid the false appearance that larger turbines kill more birds'. If total number of fatalities at a site is the variable of interest, then *their metric* is *in*appropriate. To compare deaths as a function of turbine size, then fatalities per turbine *(e.g., grouped by size categories)* is an appropriate metric and does not give a 'false appearance'. *I now understand that by using their metric that includes megawatt production, they attempted to adjust for time of operation, i.e., it is used as a surrogate. I agree that this factor should be considered in one is trying to identify which turbine types are more dangerous, but use of MW production is only a partial surrogate for operation time (which they admit in their response). So, for example, if 100 birds are killed by 100 turbines of each size class operated the same length of time, I would conclude that bird mortality is equivalent between the 2 turbine sizes. However, if one accounts for MW production, the larger turbines would have a smaller metric of mortalities per megawatt per year* *assuming that larger turbines generate more power. Based on this scenario, I can only conclude that on a cost-benefit basis, the larger turbines are better, but in an absolute sense of fatalities, larger turbines are no better than smaller ones.* By incorporating the MW produced by each turbine, they have simply factored in the benefit of generated power in this cost representation. *Wind speed also underlies in the MW production number, thus it is difficult to isolate tower size effects with any metric unless one can control the other factors (operation time, wind speed, etc.). I hope this clarifies my initial point that the metric to be used depends on the objective of the measurement itself.* There are advantages of using this metric, many of which are stated in appendix A, but the advantage depends on how the metric is to be used.

The authors state that at least 3 years of carcass searches are needed before stabilization of the percentage of non-zero mortality values. Are they saying that if a sample of 100 turbines is surveyed, at least 3 years are needed to estimate the percentage of those 100 that kill at least one bird? I am not convinced this is a useful metric. Rather than focusing on the turbines where zero, or even an occasional mortality occurs, should not the focus be on those characteristics at turbines where numerous mortalities occur (e.g., see figure 3-4). They proceed to interpret this result by stating that one must survey at least 3 years before getting a 'good' estimate of mortality rate. Mortality rate (expressed by fatalities per turbine per year) is not the same metric as percentage of turbines with at least one fatality. While I agree that more data is better for estimation in general, they have not demonstrated 3 years of data are necessary for a 'good' estimate of the mortality rate. The term 'good' in the context of bias would require knowledge of true mortality rate. The term 'good' in the context of precision would require some definition of what precision is needed for the estimates to be useful. *The latter can be examined with their data for those turbines searched multiple years. In other words, they could report the variation in estimates from one year to the next to estimate how many years of data are needed to achieve estimates within a certain margin of error with a stated level of confidence. In other words, given variance estimates, sample size estimation is possible.*

I disagree with the general statement that their test results for associations were statistically and biologically sound. Reasons for my contrary opinion are provided throughout my review. The predictive power of their 'models' reported in this section and later is likely a poor indicator of model validation given that the same data used to construct the models has been used to test them.

Chapter 1 Understanding the Problem

General comments

The first chapter gives an overview of the project, the objectives and their general approach toward meeting those objectives. Important definitions are made regarding their usage of terms like susceptibility, vulnerability, etc. For instance, vulnerability is measured here on a relative, not absolute basis. *The clarity provided with these definitions is helpful to the reader. However, many other terms are not as clear.* For instance, how is *bird* use measured *(I previously used the term 'habitat use')*? -this is a very important question when interpreting results. How close does a bird have to be to the reference point (e.g., rotor) to count as use? The authors use the word 'nearby' wind turbines, but I am uncertain what that implies. Is flying over an area for a few seconds treated the same as when a bird perches or hunts in the same area over several minutes or hours? Their phrasing suggests they consider the proportion of sampling periods in which use was detected, but this does not indicate duration of use per se. How do they treat observations of multiple birds at the same time? Are pairs treated as one observation? *Smallwood and Thelander responded by stating this chapter was meant to be conceptual and as such, perhaps these comments are premature.* Many of these questions are explicitly answered in later chapters, but some are not.

Specific comments
The authors correctly state that a preferred study design would have been to use a before-after control impact design but that such was not possible given the prior existence of the turbines. On the bottom of page 9, the authors present a 'model' for vulnerability as the ratio of observed and expected use. I suggest they restate this as a metric, not a model. It is not clear why the Chi-square symbols are in the numerator and denominator of this expression. A Chi-square statistic can be computed based on the sum of squared differences of observed and expected values, divided by the expected values. The authors should clearly state this is a 'goodness of fit' approach to testing. Section 1.1.3 is nice section on the difficulty of measuring impact. I would like to know, however, how the number of mortalities per year in the APWRA compares to other hazards, such as collisions with vehicles or airplanes, or deaths due to poaching or contaminants. This would give the reader some perspective on the magnitude of impacts of strike mortalities in the APRWA. I realize that for some species, e.g., the golden eagle, car collisions are unlikely, but what about other human-induced sources of mortality? *I only mention this in case the value of wind farms was being questioned (relative to its costs). Similar considerations have been ongoing, for instance, regarding the utility of dams versus the impacts they have on river systems.* Section 1.1.3, introduces the notion that by comparing observed and expected frequencies, one is able to identify which environmental factors might have a causal relationship (see p. 12, 4th and 5th sentences of first paragraph). The term 'might have' is important, because this is merely an observational study and thus causation cannot be established. The objectives are then described; the sampled population (initially) is identified as approximately 28% of the APWRA's turbine population due to limitations placed on access, and a brief description of 'midcourse' corrections is stated. Section 1.1.4 introduces the idea of 'use versus availability' in terms of assessing mortalities and associations with turbine location by considering what percent of mortalities one would expect given random use of the sampled area versus the number actually observed. This reasoning is the basis for much of the statistical testing (Chi-square goodness of fit tests) presented later in the report. I question to what population is the statistical inference being made with these tests. *Smallwood and Thelander state the population is the sample used in the chi-square test. I am confused by this statement. Statistical testing consists of using sample data to infer about a larger population. No testing is needed if one's interest is only the sample. They further state that for the highly significant tests, inference can be drawn to birds using the APWRA. First, the authors are misinterpreting P-values as effect size. Use of the term 'highly significant' demonstrates this misunderstanding (common to many scientists). Second, inference to the entire APWRA is not supported statistically as discussed earlier*

and certainly does not depend on the P-value from a hypothesis test. I agree that by examining the observed/expected ratios, one can describe places where more or fewer mortalities occurred than expected with random use of the sampled area. However, is it reasonable to assume that birds use landscapes randomly? *Smallwood and Thelander admit this is unreasonable to assume, but is the null condition. However, all of their statistical tests are based on this false assumption, which calls into question the value of the P-values that are reported. These are all computed assuming the random use, which is clearly false. Anderson et al. (2000) and others have referred to these as 'silly nulls'.* On page 20, the authors mention a focused study on bird behavior involving about 1500 wind turbines. Did they randomly sample these turbines from the collection of all turbines they studied? If so, then they could make inferences to the larger collection of turbines they surveyed, but again, I would suggest they resist the temptation to infer to the entire APWRA.

References

Anderson, D.R., K.P. Burnham and W.L. Thompson. 2000. Null hypothesis testing: problems, prevalence and an alternative. Journal of Wildlife Management 64:912-923.

Chapter 2 Cause and Death and Locations of Bird Carcasses in the APWRA

General comments

This chapter basically describes the mortalities found by species, cause, season, turbine model, etc. These detected mortalities form the basis for their development of associations, causal factor identification and recommendations based on their general conclusions. Thus, how the data were collected (chapter 3) become very important in interpreting what the data represent. Obviously, the fact that they searched around wind towers suggests they are predisposed to finding mortalities due to collisions than, say, due to natural predation or other factors (e.g., disease). Statistics referring to most mortalities being found near KCS-56 turbines and Bonus turbines or during certain seasons are useful in a descriptive sense, but are less useful in inferring associations unless placed into the context of search effort, availability of turbine types, local raptor abundance, and other factors likely to influence bird strike mortality numbers. My perspective is that the most useful data from this chapter are reported as the percentage of mortalities within their search radius, based on the relative number of birds found outside the search radius. Their recognition that end towers may require a search radius larger than 50m to find 90% or more of carcasses in the 'world of the turbine' is valuable for future survey efforts. Unfortunately, I did not see any attempt to estimate their detectability within their search radius. I see little value of the statistical testing in this chapter (see specific comments below).

Specific comments

At the top of page 30, they state a total of 1162 fatalities caused by collisions and by unknown causes were found. Table 2-1 identifies all 1162 fatalities as wind turbine collisions. Why are the unknowns folded into this column of the table? *Smallwood and Thelander responded by saying they assumed the unknown mortalities were due to brid strikes.* Note that 3 of these observations are bats and 42 are unknown species or group.

By assessing the ‘efficiency’ of their search radius, I assume the authors are referring to what percentage of bird strike mortalities are contained within their search area. Thus, they are actually referencing detectability and attempting to determine the validity of 100% detection rates. In most biological studies, detectability is less than 100% and estimates of detection rates are necessary for actual abundance estimation. Clearly, the observation that carcasses were found beyond a 50-m radius indicates that their mortality estimates are underestimates of true mortality. I am curious as to why the authors did not expand their search radius to lessen this bias; however, I can appreciate that a larger search area would mean considerably more search effort that logistically may not have been possible. I am curious as to what the detection rate may have been within their defined 50-m radius. The researchers could have directly estimated detection rates using various techniques (double observer, capture-recapture, removal approaches, or distance sampling if actual distances of carcasses were measured for each carcass). Instead, they related distance to carcass as a function of bird body size, wind turbine attributes, season, etc. They later state (page 49 bottom) that they were unconcerned with underestimating mortality, yet they spend much of chapter 2 examining carcass distances to assess ‘efficiency of search radius’. What *a priori* hypotheses did they have regarding bird body size and distance from turbine? Given a clear association, how is that useful for determining detection rate? I can understand how certain information such as how carcass distances are associated with turbine height or rotor speed might be used to aid in the design of future survey efforts. For instance, if taller wind turbines displace strikes over a 100-m distance, then a suggestion would be to use a larger search area when quantifying bird strikes. This purpose is stated at the bottom of the second paragraph on page 28.

The purpose of the arbitrary distinction of small and large body lengths in section 2.2 is unclear to me, as is any age classification. The analyses that followed was size-specific, but I do not understand the reason for such a partitioning. *Further, the grouping used is arbitrary (they used a natural break in the histogram of body lengths).* I also do not understand the statement that they lacked sufficient funding to factor in the slope of the hills from each wind turbine. Are they saying they could not afford a clinometer? *The authors clarify in their response that they did not measure inclination during their data collection effort, so to return to make these measurements would require more funding.* I am curious as to how Pearson’s correlation coefficient (p. 28 bottom) was calculated for assessing the linear association of carcass distances and elevation of tower base. A given tower base may have had multiple carcasses with multiple distances. Did they treat each of the carcasses as independent observations or did they compute an average distance for all carcasses at a given tower base?

In section 2.3, the authors state on page 29 bottom that most carcasses were discovered during summer and winter. Is that because more surveys were performed then, or a greater abundance of birds were present, or a lesser number of birds were present, but they used the area over a much longer duration than passing migrants do in spring and fall? The number of bird carcasses next to KCS-56 and Bonus turbines is drastically higher than all other turbine types. My question is ‘is it the turbine type that predisposes it toward more bird strikes, or are there simply more of these turbines or that they were surveyed more often or are these turbines in places were birds are more abundant as a result of some other attribute, e.g., landscape feature? *Their response to* *this question indicated there were more of these turbines and they were searched the longest. Thus, based on these data, one should not conclude these are the most dangerous types of turbines.*

The ANOVAs reported in this section demonstrate statistical detectability of differences among means of carcass distances by tower height. I am not convinced that any of these analyses are useful given the purpose of this data collection. It was my understanding that the purpose of examining bird carcass distance was to 'assess the efficiency of their search radius'. Testing for differences in mean distances in not an effective approach to determining how large search radii ought to be at different tower locations. One needs to look at the distribution of the distances and/or actually estimate detectability of bird carcasses due to strikes with wind turbines. There are 2 levels of detectability here. The first level of detectability concerns what proportion of strike victims are beyond a prescribed search radius? The authors have collected information for estimating this proportion. The sentences that state 'Our search radius included 84.7% of the carcasses of large-bodied birds (90.5% for small-bodied birds) determined to be killed by wind turbines or unknown causes' are the most informative in this section, although I would eliminate the distinction of large and small bodied birds and eliminate the unknown cause counts. Figure 2-12 is also useful here in demonstrating that the 50-m radius contained approximately 95% of carcasses in most cases (the large variances for KCS is curious, the lone (extreme) observation for the Danwin turbine is also notable). The second level of detectability pertains to within their search radius, what percentage of carcasses is found? Later in chapter 3, page 51, they state they adopted a searcher detection rate of 85% for raptors based on Orloff and Flannery (1992) and 41% for nonraptors based on two other studies. How valid these estimates are for the current study is unclear *due to potential methodological differences (personnel, survey method, time frame, etc.).*

Other concerns are 1) why was ANOVA used for the continuous explanatory variables? Regression modeling seems to be more appropriate (which they also report, but do not emphasize). *Reduction of numerical data to categorical results in information loss, loss which is avoidable.* 2) As mentioned previously, there is a potential for confounding effects given only one variable (e.g., tower height) is being examined in each analysis. For example, in the large-bodied bird section, if the 32-m towers were placed more often on sites with greater slopes, then the comparatively large 57m average distance may be a function of slope, not tower height (they recognize this shortfall on page 45 bottom, but fail to see that there are many factors, like blade speed, that should also be jointly considered in analysis). I would suggest the authors consider regression analyses that include several pertinent explanatory variables, rather than a one-variable-at-a-time analysis approach. *If only continuous variables exist, then multiple regression can be used. If both continuous and categorical explanatory variables exist,* such as rotor direction (upwind or downwind) or wind turbine location (end, gap or interior) *then analysis of covariance is appropriate. The main point here is that by only considering one variable at a time, they are blind to problems of confounded variables, and are not able to best identify which variables are most important (when considered collectively). In their response, the authors indicated they believed the one-variable-at-a-time approach forced them to examine the data more carefully than using multiple variable (they use the term multivariate) methods. This logic would indicate that more* *sophisticated models are less useful, when in fact, they are more useful, e.g., have more statistical power.* The authors do need to be aware of possible multicolinearity problems when using multiple explanatory variables and thus examining associations of the explanatory variables is recommended. I would also suggest that the authors consider the precision of the model as well as the assumptions underlying its use. If the precision is poor, or if the underlying assumptions are not met, I would not rely on the model and its estimates or predictions. To quote Michael E. Soulé, "Models are tools for thinkers, not crutches for the thoughtless". 3) The differences detected with the LSD tests that are reported are not biologically important in my opinion (e.g., the means ranged from 26m to 33m for large-bodied birds), nor are correlations meaningful when the sample correlation coefficient itself is near zero (see page 44, bottom). If the authors disagree, they need to make a case for why the differences are meaningful, that is they must identify what is a meaningful effect size.

Finally, I do not understand the last sentence in the first paragraph of the discussion (page 45). Clearly, they do have an unknown proportion of actual carcasses given that carcasses were located beyond their search radius *(they agree with this statement), yet their sentence is 'We assume that we did not find an unknown proportion of ...carcasses' Perhaps they intended to say 'We assume that we did not find all carcasses'. Clearly, my misinterpretation of this sentence remains.* Thus, their observed counts of mortality likely do not represent all strike mortalities over the defined spatial and temporal sampling period. They later clearly state this on page 49, bottom. Perhaps they did not intend to have the word 'not' in this sentence. The authors appropriately recognize alternative reasons why bird distances may be identified farther away for turbines at the ends of strings and on hills.

Chapter 3 Bird Mortality in the Altamont Pass Wind Resource Area

General comments

In this section, the sampling methods for counting bird carcasses are described. A staggered entry of turbines was used as access to turbines became available. A second set of wind turbines was added in November of 2002, and were surveyed until May 2003. According to a statement in the executive summary, mortality estimates should not be deemed reliable until 3 years of surveying has been conducted, use of this portion of data would seem to be inconsistent with this statement. *At many places in the report, the authors have qualified results with preceding statements such as 'we have low confidence in the mortality estimates' or 'note that the correlation was low'. But having done so, they then proceed to draw conclusions as if their results are valid and/or meaningful. In their response, the authors claim their suspicions about bias in turbine access issues was warranted.- A conclusion based on low-confidence data. This approach can be commonly found in the scientific literature, but that does not make it tenable. The authors state the assumptions upon which their results are based are invalid, and then later proceed to make conclusions as if the assumption violations did not exists and the data were confirmatory. I believe the data should be reported, but suggest the authors refrain from drawing any conclusions.* For those turbines searched one year or more, temporal coverage is stated to be approximately 7 times per year. Two people searched for carcasses within 50m of each turbine and 50m beyond the end turbine. The authors state they did not estimate searcher detection and scavenger removal rates because they were unconcerned with underestimating mortality, yet later, they adopt corrective measures for these processes from other studies in estimating mortality *(I deleted the end of this sentence stating this was contrary to their prior stated indifference).* They conclude by stating their mortality estimates might be conservative. However, I would suggest they may also be overestimates if the set of adjustments *for scavenger removal and searcher detection they used from other studies* were not applicable to their surveys. *The authors claim their adjustments were conservative, but I am not convinced that estimates from other studies are applicable to their methods. For instance, if they surveyed more frequently than other studies, scavenger removal rates would be lower. If they were more diligent in surveying for carcasses, then detectability might be higher in their work.* An abundance of statistical tests were performed, testing for time variation in mortality (Tables 3-3 through 3-8). In conducting several hypothesis tests using a Type I error rate of $\alpha = 0.05$ (comparison-wise error rate), the authors are likely to have some null hypotheses rejected due to type I error. That is to say that some null hypotheses will be rejected even though they are true because the type I error rate for the entire collection of tests (experiment-wise error rate) will be much greater than 0.05. *If the authors are concerned about false rejections, they could use a Bonferroni adjustment for the alpha level for their test (which lowers the comparison-wise alpha level as a function of the number of tests performed). In doing so, the probability of at least one type I error is reduced, but the observed P-value must still be lower than the comparison wise alpha level to reject Ho. In essence, a more restrictive criterion is being used to reject Ho. Simply having the P-values listed does not take care of the problem unless they clearly state what the comparison-wise alpha level is.*

Specific comments

Their metric for reporting bird mortality is clearly the number of fatalities per megawatt of power per year. The authors give previous mortality estimates from other authors, but these were reported in deaths per turbine per year. Hence the numbers from this study cannot be directly compared unless one knows the megawatts per turbine from other studies. So, I am a bit confused by the statement (page 47) that their purpose was to estimate mortality so that comparisons could be made to other sites. (I see later in table 3-12, their use of fatalities/turbine/year for these comparisons.) *The authors responded to this by stating that almost all reports now use their metric. I hope that its limitations are obvious to all (see earlier discussion regarding this metric on page 8).* I would also be interested in knowing if their survey methodology differed from previous work. If so, then they should be cautious in making comparisons of observed mortality rates. For example, if their search methods were more thorough, then observed mortality differences may be due to detection differences, rather than actual mortality differences. *The authors agree that caution is warranted, but despite that, I wonder if they proceed without this caution in mind.*

In this same section, the authors state that they extrapolate to the portion of the APWRA not sampled in order to characterize the range of likely project impacts…. I would caution the authors that because the sample of sites they have studied is not a random sample, such extrapolation is not supported from a statistical inference perspective. The authors are aware of problems associated with nonrandom sampling of other studies (see page 179, second paragraph), but have overlooked application to their study. As stated before, if the authors can justify that the areas they studied are similar to those not studied, then perhaps one might accept some plausibility in their projections for the entire APWRA. However, they authors later state that they do not know the attributes of the tower locations not surveyed, which compels me to suggest that the authors refrain from broader inference and report their results for the observed sample of sites. Their descriptive information for the sites sampled is clearly valuable and represents the majority of the study area, thus they should refrain from making broader inferences that are not defendable. In their conclusions in chapter 9, they

On page 47, they state they were unable to search all turbine strings throughout the study or equally in frequency, so that time spans and seasonal representation varied at turbine strings. Again, I cannot blame them for logistical constraints, but they must take care in analyzing and interpreting patterns in data in light of the fact that they do not have a well-designed study in which all combinations of factors are represented with replication. *I acknowledge that the authors provided cautionary statements, but as previously stated, wonder if these concerns are forgotten when drawing conclusions.* On the bottom of page 48, in determining time since death, how much do weather conditions affect these estimates? On page 51, the authors describe adjustments they made to their observed mortality counts. For instance, they state they adopted a searcher detection rate of 85% for raptors based on Orloff and Flannery (1992) and 41% for nonraptors based on two other studies. They further assumed scavenger removal rates (differential by small and large birds) based on Erickson et al. (2003). They added 10% to these rates for the second set of wind turbines. How valid any of these estimates are for the current study is unclear. Thus, I am skeptical of their estimated number of fatalities for the APWRA given in the executive summary and presented in tables 3-10 through 3-12. Clearly, differences in personnel, search effort and design, habitat, weather conditions, predator density, etc., would lead to differences of these rates from other studies. I suggest the authors concentrate on the observed mortalities in their study and consider optimal strategies for harnessing the information contained therein. As an example, Figures 3-5 through 3-14 present means and standard errors of mortality estimates (per MW). The use of standard errors implies the authors are inferring to a larger population mean. Therefore, it is important to clarify to what population their interval estimates refer. If they are estimating the mean for the entire APWRA, then the same limitations of inference apply as stated before, given the nonrandom collection of sites that have been surveyed. *In their response, the authors state the unsampled turbines did not differ from the turbines they searched. Is that true for all aspects, e.g., landscape features, bird abundance and use of those areas? I have no reason to believe that they purposely biased the selection of turbines, but I would be interested to know how they can guarantee that these unsurveyed locations were similar in all aspects to the measured turbines.*

If, on the other hand, they are estimating mortalities at surveyed sites for the entire year based on a sample over time, then how they sampled over time becomes important in considering the validity of the inference being made. I suggest the authors consider using descriptive statistics of mean and standard deviation (not standard error), which are informative in terms of describing their sample of sites surveyed. This data shows that higher mortality occurred for red-tailed hawks and barn owls during the year

1999/2000, although the reasons for this are not stated. There was also more variability in observed mortality rates during this year. In some cases, consideration of the variability is just as interesting as a measure of center, so they might consider why there was more variation in mortalities this year. By relying on descriptive statistics for their sample of sites, the hypothesis testing in tables 3-3 through 3-8 is not needed. *To their credit, the authors seem very willing to make many of the suggested changes if allowed the opportunity to revise the report.*

In their discussion in section 3.4, the authors reference higher mortality rates at Sea-West-owned turbines than other portions of the APWRA. I was not able to find the supporting evidence for this statement. In addition, I fail to find how ownership would affect mortality per se. Perhaps the ownership issue is tied to some other attribute of which I am unaware. I would like to see more clarification from the authors about the point being made. Further clarification regarding the biological significance of any raptor mortality (last sentence of page 76) would be beneficial. *The authors provided further clarification in their response.*

Chapter 4 Impacts to Birds Caused by Wind Energy Generation

General and specific comments

Examination of bird mortality in relation to bird abundance is a worthwhile endeavor; however, actual bird abundances were unknown. The authors have relied upon measures of relative abundance, determined from point counts. Usefulness of such indices relies on the assumption that the detection rates of birds are similar across time and space. The assumptions should be explicitly stated. There are a host of reasons why this assumption is likely to fail, including observer differences, animal differences and environmental differences (see Anderson 2001, Ellingson and Lukacs 2003). Logically, I would anticipate some sort of positive association between abundance and mortalities, however the positive associations estimated in figures 4-1 and 4-2 are likely not valid for the above detailed reasons. Figure 4-5 B actually demonstrates a trend counter to this assumed association, so an explanation would be helpful for the reader.

Comparison of mortality rates reported from several previous studies is unwise given the vast differences in survey designs, methodologies, site-specific and surveyor-specific differences. I appreciate the attempt to synthesize information from several studies, thus placing their results in context. However, almost all of the assumptions/adjustments made in this section are likely to be false and the extent of bias due to this failure unknown.

In the results, the authors state that bird mortality did not correlate with radius of search around the wind turbine. I am likely misinterpreting this statement, but if one increases a search area, one can only find more carcasses, not less, so I find this confusing. The data presented on mortality at APWRA from 1988 to 2000 has not been described in terms of consistency of methodology, surveyor ability, effort, consistency of environmental conditions that might affect detection rates of carcasses, wind turbine numbers, etc. Thus, I find it difficult to accept the reported trends as meaningful. The authors recognize the importance of standardized methods (see page 86), yet they have not made it clear they have met this requirement in their analyses and in fact, they explicitly state some differences among these studies.

References:

Anderson 2001. The need to get the basics right in wildlife field studies. Wildlife Society Bulletin 29:1294-1297.

Ellingson, A.R. and P.M. Lukacs. 2003. Improving methods for regional landbird monitoring: a reply to Hutto and Young. Wildlife Society Bulletin 31: 896-902.

Chapter 5 Range Management and Ecological Relationships in the APWRA

General comments

Methods described here are not detailed enough to fully evaluate this section. For instance, did they randomly place their transects (string and grass) within a defined area around the turbine string, or were these haphazardly or judgmentally placed? How was average vegetation height measured along a transect, based on every plant or at specific points (i.e., a point transect sampling approach)? How might detectability of cattle pats, rabbit pellets, lizards, mammals, etc., differ in different locations? As before, to what population is inference being made with the statistical tests? *The authors' response to this is that it depends on the strength of the test or how small the P-value was. Identification of the population to which a test or inference is being made does not depend on a P-value. This clearly is a misconception of what a P-value represents. There has been a tremendous amount of discussion on the utility of hypothesis testing, what P-values represent, etc., in the biological literature in recent years (see, for example, Cherry 1998, Johnson 1999, Anderson et al. 2000). I suggest the authors read some of these articles as well as counter viewpoints (Eberhardt 2003).* Use of the word 'significant' needs to be clarified as statistically detectable rather than biologically meaningful. I recommend the authors reserve the word significant only when referring to a biologically meaningful observation. Their one-variable-at-a-time approach may confound the observed associations. For instance, the authors give many examples of how vegetation height differed according to aspect, physical relief, etc. Although I am not convinced these are meaningful differences, they do indicate numerous variables are being considered, and a one-variable-at-a-time analysis procedure has inferential limitations which have been discussed previously. What is the rationale for associating turbine or tower type to lizard counts?

Specific comments

Use of the phrase, 'tended to be significant' on the middle of page 91 is either improper interpretation of P-value as related to effect size or from a decision making perspective of null hypothesis testing a way of circumventing a yes-no answer in the formal test. First, a P-value is not an indication of effect size, after all, its value can be changed simply by changing the sample size; the same estimate can yield differing P-values. A P-value can be interpreted as the probability of observing a result as or more extreme than that observed given the null hypothesis (Ho) is true.

The biological significance of the observations has not been justified, conclusions have been based on statistical detectability. For instance, the mean difference of vegetation height comparing heavy versus intermittently employed rodenticides was 4.28

cm). Is this a meaningful difference, i.e., *do prey species and/or raptors perceive this difference? The authors responded to this question by stating that published literature indicates changes in grassland use by animals according to vegetation height, but I still have no idea if a difference a few centimeters falls within that umbrella statement.* The summary table 5-25 summarizes their findings, but underlying all of these associations are questionable interpretations of biological importance.

The mean differences in Table 5-1 are confusing, for instance, when comparing plateau to plateau (these are the same variable) and when comparing plateau to peak and slope (one cannot perform LSD on 2 *different* variables), they must use another multiple comparison procedure which allows combinations of means. Note also that it is improper to follow nonrejection of an ANOVA with LSD multiple comparisons because there is no control for type I error in such a case. *Apparently, I did not note that their alpha level was 0.10 in this chapter. It is unclear why they used a more liberal test in this chapter, which results in a higher Type I error rate.*

Table 5-3 gives several correlation coefficients, most of which indicate a weak linear association at best, yet they highlight the statistical detectability of these cases. For example, the last sentence in the discussion on page 54 restates that vegetation height correlated positively with number of cattle pats, but the sample correlation coefficient was only 0.19, a weak association at best. *Smallwood and Thelander state that they did not characterize them as significant. By simply reporting these values in the results section, I agree. However, when they reiterate a positive relationship in the discussion, they are then interpreting the result as significant. By not having the qualifier 'weakly' positively correlated in their discussion, the reader may be misled to think the correlation is relevant.* One correlation coefficient given for vegetation height and 'percent in canyon' is moderate (r = 0.46), but I am not certain this is meaningful. The population of turbines being measured is clearly stated as the 1526 wind turbines measured through August 2002. Thus, statistical inferences might be made to this collection of turbines if sampled appropriately *(they agree)*. The important consideration then becomes how the turbines were sampled. Were transects randomly placed within a defined area around each turbine or the turbine string? If a multistage random sampling design were implemented, estimates would be possible at both the individual turbine and turbine string level. My impression is that random placement was not made at either scale, thus limiting their inferential capability. In my comments on this section, I have provided a few examples in which I question the value in the perceived associations. Rather than repeat the same concerns for each table/response variable combination, consider my concerns to apply to all of the variables and associated analyses performed in this chapter (cattle pat counts, cottontail index, lizard index).

References

Anderson, D.R., K.P. Burnham and W.L. Thompson. 2000. Null hypothesis testing: problems, prevalence and an alternative. Journal of Wildlife Management 64:912-923.

Cherry, S. 1998 Statistical tests in publications of The Wildlife Society. Wildlife Society Bulletin. 26:947-953.

Eberhardt, L.L. 2003. What should we do about hypothesis testing? Journal of Wildlife Management 67: 241-247.

Johnson, D.H. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management 63:763-772.

Chapter 6 Distribution and Abundance of Fossorial Animal Burrows in the APWRA and the Effects of Rodent Control on Bird Mortality.

General comments

The idea that reducing raptor prey populations in the APWRA may discourage raptors from visiting the APWRA makes intuitive sense, although I am not sure if this effect can be achieved at such a large scale (as opposed to more locally near wind turbines). I am not convinced that elimination of one component of prey, specifically ground squirrels, would achieve the desired result either. I commend the authors for making observations regarding gopher and squirrel burrows and their proximity to turbines, and developing research regarding the variables. Their objectives are clearly stated as relating ground squirrel and pocket gopher distribution and abundance to rodent control intensity, physiographic and turbine attributes, and comparing raptor mortality to densities and contagion of burrow systems, but a priori hypotheses are not explicitly stated. Burrow densities are implicitly being used as indices to abundance. I have no knowledge of whether or not this assumption is reasonable. For example, the burrows may represent a population size that existed several years prior to the current observed raptor morality or numbers of burrow per animal may differ depending on landscape or predator abundance features. How ephemeral are these burrow systems? The seasonal effects reported in setion 6.3.2 indicate tremendous burrow variability within a year, but are numbers of individuals fluctuating that much? *The authors provided many clear answers to these questions in their response.* Changes in the numbers shown in various figures make me question the relevance of burrows as an index to prey abundance, let alone prey availability. See my discussion on chapter 4 regarding the utility of indices. *Here, I was referring to the tremendous fluctuations in burrow systems as seen in their figures. Smallwood and Thelander state in their response that these fluctuations are real (well established in the literature).* Two metrics of contagion used in their analyses. Given the observed variation in burrow numbers throughout a year, how did they relate observed bird mortality over a year or several years to burrow contagion? *Here, I was asking what measure of contagion did they use for a given year if the measure changed throughout the year.*

Specific comments

Wind turbine strings studied were selected arbitrarily (not randomly), hence limiting statistical inferences *(the authors agree and clarify that this was exploratory work))*.

Figure 6-4 (and later 6-45 and 6-46) presents results comparing burrow system densities between areas with rodent control and areas lacking rodent control, however, for the latter, only 3 observations were available. The strength of evidence here is very weak *(the authors agree)*. In figure 6-5, they discarded an outlier without explanation. If a poor measurement was made so that this result was unreliable, then state this clearly. If, however, it does not fit their predefined opinion on what should occur, then this is not a viable reason to omit it from the analysis.

The authors resort to transformations at various places without explanation or consideration of what is then actually being compared. For example, what does Figure 6-6A tell us? The variable based on a log-log regression in this figure and Figure 6-7 has not been explained. *A description of these figures is provided in the text, but the authors did not state why a log-log plot is used rather than working with the data on the original scale.* Why are normal curves shown in all frequency distributions? *There was no meaningful reason for doing so.*

I find it curious that ground squirrel avoidance was stated to differ between summer and other seasons (see page 124), given the degree of overlap of *summer with fall and spring* (Figure 6-27B). *The authors clarify in their response the difference is detectable between summer and winter.*

The term 'tended to differ significantly' or 'tended to correlate' is used at various places (e.g., see pages 124, 146, 166) to reflect a P-value less than 0.10 but greater than their stated alpha level of 0.05. As previously stated, this is either improper interpretation of P-value as related to effect size or from a decision making perspective of null hypothesis testing, a way of circumventing a yes-no answer in the formal test. Note again that it is improper to follow nonrejection of an ANOVA with LSD multiple comparisons (see page 124, first paragraph of section 6.3.3) because there is no control for type I error in such a case.

The reported correlations that follow are weak to moderate at best and thus I would not read too much into the statistical detectability.

From a nontechnical perspective, I find the observed relationship between rodent control intensity and pocket gopher burrow density interesting in that its highest level is at a moderate level of rodent control. Cottontail burrows demonstrated an opposite pattern. Why would gopher clustering differ by aspect for control areas, but not for nontreated areas? I could not find explanations for several reported effects.

Comparisons of raptor mortality and small mammal burrow distributions were executed only considering burrow density and thus are prone to many confounding effects as previously stated. At various times, the authors recognize the potential complexity of what they are attempting to measure, but they then put this consideration aside and proceed to analyze, interpret and conclude. *They respond by saying what else are we to do? My suggestion would be to eliminate material where treatment effects cannot be isolated, or at the very least, only report descriptive information and do not proceed to interpret it. In other words, if one cannot isolate a treatment effect, or cannot meet important assumptions, than no modeling is better than proceeding as if everything was valid.*

The finding that mallard mortality was related to rodent control intensity was dismissed as a spurious effect. I agree with this conclusion, but it illustrates an important point. When one can develop an ecological explanation for an observed result *a posteriori*, the result is more heavily weighted as 'truth'. I believe this approach to science is ubiquitous, but not ideal and has lead to many spurious results. *Their response to this issue was commendable.*

Chapter 7 Bird Fatality Associations and Predictive Models for the APWRA

General comments

The authors begin this chapter by stating the importance of identifying causal factors of bird fatalities. They then state that collisions are rarely observed and that inferences must be drawn from carcass locations. Such inferences are merely associative, not causative. I suggest the authors present their observations in the former context as I believe the latter is not attainable within the context of the current study. Causation is best established through experimentation and requires reasoning beyond statistical analysis. Establishing causation with observational studies requires meta-replication of the system of interest, consistency of results, and a plausible explanation for the observed behavior. Preferably, potential hypotheses should be formulated *a priori* and the evidence for each hypothesis should be documented with the observed data. Far too often, ecological explanations are developed *a posteriori*, which leads to spurious results. The authors are aware of possible spurious results (see page 218 discussion of mallard fatalities), but they fail to see this potential when explanations can be developed *a posteriori*. *Here, I was referencing their willingness to describe the mallard fatalities as spurious, but they did not warn the reader (to my knowledge) that other detectable associations might also be spurious.* More importantly, they fail to see this potential in their overall approach to analysis. For instance, 34 explanatory variables have been measured at each turbine site, 12 bird species examined leading to hundreds of single variable tests of associations (Tables 7-1, 7-2, 7-3). There is great potential for Type I error (rejection of the null hypothesis of no association when there is no association) when considering all of the tests being performed. A total of 408 tests were performed in Tables 7-1 and 7-2. The probability of not making a Type I error (using the stated $\alpha = 0.05$ level) is $(0.95)^{408}$. Thus, the probability of making at least one type I error is $1-(0.95)^{408}$, which is essentially one. *In their response, Smallwood and Thelander seemed to have confused the alpha level of a test and a P-value; it is as if they want to vary the alpha level according to the observed P-value. The alpha level is a preset measure that is decided prior to data analysis based on the level of Type I error that is acceptable to the researcher. I have not overestimated the probability of at least one Type I error. The P-value is entirely data-dependent. Note also that probability stated above is for at least one type I error, therefore this does not mean that out of 408 tests, only one will be a Type I error. The authors would like to act as if only one of the 408 tests was a false rejection; this is a misinterpretation of statistical hypothesis testing and surprises me.*

A simple approach to reducing overall (experiment-wise) type I error rate (α_e) is to lower the comparison-wise error rate (α_c), such that $\alpha_c = \alpha_e/2k$ where k is the number of tests being performed, i.e., a Bonferroni procedure. *Thus, to maintain and experiment-wise error of 0.05, the authors should use a comparison-wise alpha level of 0.000006, in which case the authors will be surprised to find that even their results with P-values of 0.001 would not be rejected. Such a small alpha illustrates the severity of the situation; it is not intended for actual use. The authors need to reduce the total number of tests performed (which reiterates why the one-variable-at-a-time approach is not recommended) and then use a Bonferroni correction. Even then, the comparison-wise error rate will be small, and thus, many reported detectable results will likely not be detectable.*

The authors have counted fatalities at all locations surveyed and measured variables (both environmental and turbine-specific) in an attempt to reveal 'robust' patterns. It is not clear to what the intended patterns are 'robust'. Based on the observed patterns, predictive models were developed, yet the models they developed were never clear to me. For instance, one could develop a model such that the expected number of fatalities Y (per MW per year) is a function of explanatory variables X, W and Z ($Y = \beta o + \beta_1 X + \beta_2 W + \beta_3 Z$). Different modeling approaches have different link functions, for example, Poisson regression typically uses a log link function (which is different than a log transformation). Alternatively, one could use logistic regression which uses a logistic link function to model presence/absence of mortality as a function of the explanatory variables. Model selection, that is identifying the best or set of better models in the set of possible models would be an important component of such an endeavor as well as assessing the goodness of fit of the most general model in the model set. Burnham and Anderson (2002) is one of many pertinent references on the subject. The main point here is that I do not really understand what their 'model' is. *The authors' response is that their models were summations of accountable mortality across the variables selected for the model. First, the one-variable-at-a-time approach is wrought with problems as previously discussed, including confounding. Therefore, anything based on a sum of these variables has inherent limitations. Second, the assumption of additivity of these variables is questionable, in part because their one-variable-at-a-time approaches have problems of confounding and also because variables may have interactive effects. Third, I am not convinced they are describing a model; rather it sounds more like a metric of sorts. A predictive model might look like the following example using multiple regression:* $E(Y) = \alpha + \beta_1 canyon + \beta_2 towerheight + \beta_3 turbinedensity + \ldots + error$ *where Y represents the number of bird strikes. Such a model considers several explanatory variables simultaneously in the estimation process as opposed to one-variable-at-a-time approaches. Note that the variable 'canyon' is an indicator variable, so this is really an example of analysis of covariance.* Is the assumption that their predictive 'models' are relatively precise appropriate (page 222)? Testing of their 'models' appears to have been performed using the data to develop the model, which is an inappropriate means of evaluating models (see Olden et al. 2002). The authors should consider using a portion of their data for model development, reserving the other portion for model evaluation.

Specific comments

On pages 179 to 182, the authors describe an abundance of previous studies which have presented conflicting conclusions regarding causal factors of collisions. Such disparity suggests to me, as stated earlier, that demonstrating causation is not a simple task and the actual mechanism underlying bird strikes may be very complex, e.g., a combination of many environmental and turbine-based factors. Again, the overall analysis approach has been to look at associations one explanatory variable at a time, thus leading to potential confounding effects and spurious results. *For clarification, I am not suggesting multivariate statistics be used (for example, MANOVA or factor analysis), I am stating that multiple explanatory variables be simultaneously used in modeling, in which case model selection methods will play an important role. Use of multiple explanatory variables is still considering a univariate modeling approach, i.e., there is one response variable (mortalities per unit of search effort).*

On page 182, last full paragraph, the last sentence is unclear. I believe the authors meant to imply that some turbine attributes were collinear or highly correlated, thus similar associations with bird mortality were observed when looking at each variable separately.

On page 184, the first three paragraphs of section 7.2.2 are wrought with lack of information and misstatements. The authors state that the assumptions of the corresponding hypothesis tests were satisfied, but they do not state what those assumptions are and how they were assessed. For example, Pearson's correlation assumes bivariate normality. Did they assess normality of each variable? How? *(Their response is essentially 'we did not, but other biologists do not either'.) Such reasoning is for lack of a better word, ridiculous. First, normality is easily assessed and if violated, would suggest use of Spearman's rank correlation. This is a basic concept that is taught in introductory statistics classes.* The least squares regression models they used assume the errors are independently and identically distributed as a normal distribution with mean zero and constant variance. How did they assess normality of the residuals? How did they test equal variance? Analysis of variance (ANOVA; which is really regression with a categorical explanatory variable) has the same assumptions. Were these also assessed for these analyses? *Their lack of attention to important assumptions leads me to conclude that extensive consultation with a statistician is needed.*

Misstatements include 'Correlation analyses are summarized by the coefficient of determination, R^2, when prediction is the ultimate objective. I believe they meant to say 'Regression analyses….' The coefficient of determination measures the percent of variation in the response variable that is explained by the regression model (i.e., the collection of explanatory variables). Correlation analysis is merely descriptive and summarizes the degree of the linear association between variables. Thus, one can have a strong association (curvilinear) but Pearson's coefficient will be small (close to zero). The statement 'We report weak and nonsignificant correlations when doing so meets our objectives.' -the latter part sounds dubious and confuses several issues. Statistical detectability is directly affected by sample size, n. Thus, with large n, it is possible to have a sample correlation coefficient of r =0.1, and yet have an associated P-value = 0.001 for the test of Ho: $\rho = 0$ vs. Ha:$\rho \neq 1$. In this case, the linear relationship is very weak and is not notable. *The authors clarify that nondetectable results can be interesting, particularly when they contradict previous conceptions, to which I agree.* Alternatively, you can have small sample sizes that result in nonrejection of Ho even though r = 0.75 indicates a fairly strong positive linear association. I agree that in the latter case, such results may be reported as long as they are taken as suggestive, not confirmatory.

Given the collection of several explanatory variables, the authors should consider using partial correlation in which one or more variables are controlled when considering the association between two variables of interest. *(I have made similar statements referring to use of multiple regression, yet in this context, the authors agree with the suggestion.)* They also incorrectly state that the coefficient of determination (R^2) is based on the steepness of the regression slope. This is only true within a specific context in which one considers a specific data set and several lines that are being fit simultaneously to the data. In general, the coefficient of determination is defined as the portion of variation in the response variable that is explained by the explanatory variable. For example, if all observations fall on the fitted regression line, then r^2 is one, and this is true regardless of the slope of the line unless the slope is zero, in which case there is no variation in the response variable. *In their response, the authors clarify they were referring to the sample value, r^2. This does not change my original point.*

They state that several key assumptions of ANOVA were not met due the absence of a block design. A block design is not necessary for ANOVA, blocks are sometimes useful for reducing variability, but their absence does not preclude assumptions from being met. Equal treatment replication (balanced design) does not preclude successful analysis via ANOVA or other techniques. However, when all treatment combinations are not represented, e.g., fractional factorial designs, then considerable thought must go in to analytical approaches for meaningful comparisons that isolate treatment effects to be made.

I do not understand what the numbers in Table 7-8 on page 223 represent. The description is that the numbers represent the largest accountable mortality values calculated from the chi square tests…. Similarly, I do not understand how a specific wind turbine attribute can be reliably associated with X% of mortalities (pages 224-241), given the combination of variables at any given turbine and the inability to control all other factors in their analysis. It follows that I am not confident that their form of model assessment (described as the percentage of correctly predicted dangerous turbines where species-specific fatalities occurred) is a useful metric for assessing model performance.

References:

Burnham, K.P., and D.R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information Theoretic Approach. 2nd ed. Springer Verlag, NY.

Olden, J.D., D.A. Jackson, and P.R. Peres-Neto. 2002. Predictive models of fish species distributions: A note on proper validation and chance predictions. Transactions of the American Fisheries Society; 131: 329-336

Chapter 8 Bird Behaviors in the Altamont Pass Wind Resource Area

General Comments

Quantifying bird behavior or intensity of use is an enormous task that is very complex in its actual measurement and analysis. My understanding of their approach is briefly given below. The researchers surveyed 61 plots encompassing between 6 and 52 wind turbines. A total of 1500 wind turbines were surveyed, although these are not independent in the sense that they are located in strings. These plots were surveyed on 4 separate occasions over a 7-month period. The plots were delineated with a 300-m buffer around the focal wind turbines. Two observers scanned the plot area over a 30 minute period, and recorded bird locations, behaviors, etc., at the turn of each minute. Basically, this means there are 30 points in time for each session. Behavioral events were also recorded, such as flight through a string of wind turbines, etc. A total of 120.6 hours of sighting was executed. Again, I commend the authors for their efforts.

From an inferential perspective, I would ask the authors to clarify how these specific sites were selected for observations. It looks as though a variety of wind turbine types were selected purposefully (Table 8-1) that cover a previously referenced area (see page 246 bottom), but a stratified sampling approach could be used to ensure random selection and representation of all turbine types as well. *The authors clarify the sample is all Set 1 turbines, thus, inference is limited to this set and should be made to APWRA.* In reference to their analysis, I am concerned that their approach is inadequate to identify meaningful relationships for reasons similar to that stated for other chapters.

First, the approach of examining one variable at a time oversimplifies what is a very complex situation. For example, by only considering minutes of perching by temperature levels, the observed difference of observed and expectations may be the result of another variable, such as wind speed, which often is associated with time of day and temperature. That is to say that such an approach does not identify causal mechanisms. To their credit, the authors do mention another source of complexity, the notion that birds may adapt their behavior in response to the presence of wind turbines (page 246). But, I am not at all confident in *m*any of their findings in this entire chapter because of their one-variable-at-a-time goodness of fit approach to analysis. *The authors have responded by stating some associations are irrefutable. For instance, they found golden eagles perched disproportionately more in canyons. I will again clarify my concern that one-variable-at-a-time approaches can be misleading. While they may have found greater than expected perching in canyons (by considering all perching observations), it may be , for example, that specific turbine types were located more often in canyon areas, and that eagles prefer to perch on these turbines. Thus, it is not the presence of a canyon that is driving the response, rather it is another variable confounded with presence of a canyon. Their conclusion that perching occurred more often in canyons is not incorrect, but it may be misleading, leading one to believe that canyons themselves induce more perching rather than another possible attribute of this study area. I am not stating that turbine types were disproportionate in canyons, but by only considering one variable at a time, they have eliminated the means for controlling for other explanatory variables.*

Second, I question whether many of the stated ‘significant differences’ are biologically meaningful. For example, in Table 8-6, the chi-square test for time of day effects on perching minutes of all raptors resulted in a P-value less than 0.005, yet the percent deviation from the expected value is less than 3 percent for all categories. In the same table using temperature as the explanatory variable for golden eagles, there is a 20 percent negative measure for temperatures of 60-69 degrees and a 21 percent positive measure for temperatures between 70 and 79 degrees. Do the authors believe that golden eagles perch less due to temperature in the 60s and more in the 70s and then less in the 80s? *Their response is that they are simply reporting all of their results. I am suggesting they be more responsible in interpreting their relevance.*

I recommend that the analysis approach in this chapter (and several others) be changed. I would need to know more to specifically advise on how they should proceed, but I will make the following statements and suggestions. First, birds (or animals in general) do not use habitat randomly. Any assumption of random use is a ‘silly null’ hypothesis which is certainly false (see Anderson et al. 2001). (*see previous discussion of this issue).* Second, while I agree that understanding how birds use APWRA would be useful in putting bird fatalities in context, I fail to see how associations with variables such as temperature would be useful. *(Here, I was referring to the inability of managers to control environmental variables, not the biological associations. Smallwood and*

Thelander have erroneously jumped to the conclusion that this statistician does not understand any biology.)

Third, I suggest the authors condense the 30-minute level information to percentages for that survey period and consider relating the percentage of time perching to the set of meaningful explanatory variables collectively. This approach treats each 30 minute period as the measure made on each sampling unit (plot). Such an approach eliminates concerns about the covariance structure between successive minute-by-minute observations. Whether or not one needs to consider the relationship between survey periods on the same site is another issue (perhaps repeated measures structure should be used?). Once a reasonable modeling approach is identified, they must develop appropriate models for consideration and use a well-defined model selection process.

Specific Comments

For me to completely comprehend the intricacies of the data collected would require considerable face-to-face interaction with the researchers. Examples of my uncertainties are stated below. I am not sure whether the focal set of wind turbines was always a complete string Also, based on figure 8-2, it looks as though a 300-m buffer may include additional sets of wind turbines; thus, they too provide opportunities for perching, 'dangerous flights', etc. How is their presence handled in terms of analysis at the wind turbine level even though they are not the focal set? I failed to see the distinction between plot level and string level of analysis mentioned on the middle of page 247. *Their response indicates all wind turbines in the observation plot were considered.* How do they count number of minutes perched if 2 birds are perched in the study area for the first 5 minutes of the study period. Is that 10 bird-perching minutes? If so, then the number of bird-minutes is the metric, rather than minutes alone. They state on page 247 that for each record, they recorded the species, … predominant flight behavior, flight direction, distance to nearest turbine, number of passes by a turbine, and flight height relative to the rotor zone. How does one record flight direction if they traveled a flight was multidirectional, circular, etc? Is distance to nearest turbine the smallest distance observed during the entire flight? How does one define a pass by a turbine? If turbines are in a string a bird flies 100 m above the string, are all of these turbines counted? How much error is there in measuring flight height relative to rotor zone when birds are considerable distance from the observers? *The authors offer useful responses to these questions. Note the importance of many of these details. Two birds perched for 5 minutes is treated the same as 1 bird perched for 10 minutes, yet the former situation has more birds susceptible to strikes than the latter.*

Birds may have exited the area for more than 30 seconds, only to return again and be considered as a new bird. Pseudoreplication is a concern here, but clearly the researchers cannot be expected to recognize individual birds per se. Several quantitative or ratio level variables, such as temperature and wind force, were reduced to ordinal categories. Information loss occurs in such a process and is unnecessary. I suggest the authors use these variables as continuous explanatory variables in a model effort other than chi-square goodness of fit tests. I do not understand why they compared the correlation of flight frequency in the rotor zone with flight time to that of perching time (page 256). I would assume that when a bird is perching, it is not flying at all, and thus is not flying in the rotor zone. I also do not understand the interpretation of frequency of behavioral observations during a 30-minute session on page 256. If the initial presence of observers is modifying bird behavior, then this is an important observation, which suggests that an initial 'settling' period should occur prior to the actual observation period. However, I am not sure this was the point they were trying to make. Finally, it is hard to evaluate the plausibility of their discussion points and recommendations at the chapter's end given the lack of enthusiasm I have for their analysis methods. I will say again that they do have valuable information, some of which was presented descriptively, that should be considered as meaningful.

References:

Anderson, D.R., W.A. Link, D.H. Johnson, and K.P. Burnham. 2001. Suggestions for presenting the results of data analyses. Journal of Wildlife Management 64:912-923.

Chapter 9 Conclusions and Recommendations

General comments

This section of the report was clearly written and suggests several management alternatives. It is my belief that some form of adaptive management (Walters 1986, Walters and Holling 1990) should take place given the amount of data already collected by this project and others cited therein. My opinion is there is still uncertainty regarding the causal mechanisms behind bird strike mortalities. However, I think it would be feasible to evaluate the effects of many of the recommendations made in this chapter in a cost-effective manner. For example, I assume it would cost very little to paint blades for a sample of turbines. Assuming that experiments can be conducted that control for other potential factors affecting mortality, one can isolate the effect of the management action.

Specific comments

Their first recommendation, at least on the surface, seems reasonable. If sampling in the WRRS program is haphazard and/or voluntary-response based, then from a scientific monitoring perspective, the data collected is of little value. That is not to say that there is nothing to be gained by observations of maintenance workers in the area, because the cost is presumably nothing. I am unsure that the comparisons of observed fatalities are fair (same time period, same locations surveyed, etc.), but if they are, and a consistent relationship between the 2 methods could be established, then such a system would be similar in worth for trend estimation. However, consistency is something not easily obtained in any index-based study (Anderson 2001).

Their conclusion regarding rodent control is counterintuitive to me, but that does not make it wrong. However, I question whether the conclusion is correct given the potential weaknesses associated with assessing the effects of rodent control treatments (see chapter 6 evaluation). They further state that even if rodent control were effective, displacing raptors would result in a net loss of raptors from the remaining habitat. *Smallwood and Thelander state that populations would be reduced through displacement because these species cannot be crowded into smaller spaces.* That is only true if populations elsewhere are at carrying capacity, which I doubted to be the case. *They state that carrying capacity is difficult to define, yet this assumption is implicit in their above statement regarding crowding into smaller spaces. Secondly, would not displacing birds from the APWRA into other suitable habitats reduce the potential for bird strikes, a likely* *goal of everyone involved? (I deleted a sentence here that stated if birds were at carrying capacity there might be less concern about the observed mortalities)*

Their third recommendation (and its subcomponents) is similar in the underlying idea to the second recommendation: reducing prey availability may reduce raptor susceptibility and thus fatalities. Thus, I find it interesting that they advocate ceasing the rodent control program, but encouraging fossorial animals to be farther from wind turbines. *Note that I referred to similarity in the underlying idea of reducing prey availability, not the spatial differences of these approaches that they refer to in their response.* However, I agree that one might want to eliminate the rodent control program for other reasons (e.g., adverse impacts on other species of importance).

In their test of perch guard effectiveness (recommendation #4), did they control for other factors that may affect mortality? Similar to the results for rodent control, it may not be the method itself that is lacking effect; it may be the implementation, for instance, if the chicken wire readily falls apart. *I would think that any effort to keep birds away from the towers at all times might reduce their susceptibility to strikes. So, the question remains as to whether effective perch guards could reduce strikes. The authors readily admitted that the perch guards implemented thus far were ineffective, e.g., chicken wire falling apart.*

Their conclusion that wind turbines at the end of strings are edges of clusters kill disproportionately more birds is very plausible, and hence their suggestion of adding pole structures is worthy of consideration for experimentation. However, it may be that by adding pole structures more birds will collide with the turbine because of the visual impedance mentioned in recommendation number 9. *I misinterpreted this recommendation. However, pole placement at the end of turbine strings might offer additional perching opportunities, thereby bringing birds in closer proximity to the turbines and increasing susceptibility.*

Most of the remaining recommendations are yet to be proven as effective management options. I suggest that pilot studies be used in which the efficacy of a small set of management actions (a subset of their listing) can be evaluated without the confounding effects of other possible mortality factors. *The authors respond by saying that pilot studies have been tried before and were inconclusive. Why then do the authors believe enough is known to implement mitigation measures at a very broad level?* Based on my limited knowledge from this report, I am not convinced that enough is known to warrant universal implementation of certain mitigation measures (see page 348 bottom). *I make this statement because I do not believe causal mechanisms have been identified and because of confounding issues from their one-variable-at-a-time analyses.* I also question the degree of error one would have in the estimated number of bird mortalities over 10 years as the input to Smallwood's estimator of are for support described on the top of page 348. *To clarify, the estimated number of bird mortalities over 10 years is being based on what, the number of mortalities found during their study? Extrapolating any such estimates to other periods is untenable, and as such*, unreliable input into a model almost certainly results in unreliable output.

In section 9.3, Thelander and Smallwood state they were unable to extend their model predictions to the turbines not characterized. Given that an appropriate predictive model exists (which I have previously questioned), I would suggest that to properly evaluate the model, these sites would provide an independent means of evaluating the model *via future data collection*. Their use of the same data for model development and testing violates the independence necessary for proper model evaluation (see Olden et al. 2002). I agree with much of their description of limitations in this section. Many of the concerns they state are equivalent to what I have stated in my review and provide the basis for my concern regarding the validity of stated conclusions. I assume that when they are referring to multivariate statistical methods on page 353, they are referring to multiple response variables, not multiple explanatory variables, but their one-variable-at-a-time approach to analysis indicates there may be confusion regarding this terminology. Multiple regression and analysis of covariance are considered univariate methods because one response variable is being considered. I suggest that these methods could be employed to better examine mortality as a function of several explanatory variables. Multivariate methods generally refer to analysis in which multiple response variables are being considered simultaneously, which requires consideration of the covariance structure among these variables.

I am not sure how they arrived at their estimated mortality reductions on page 354. These are likely purely speculative.

References:

Anderson 2001. The need to get the basics right in wildlife field studies. Wildlife Society Bulletin 29:1294-1297.

Olden, J.D., D.A. Jackson, and P.R. Peres-Neto. 2002. Predictive models of fish species distributions: A note on proper validation and chance predictions. Transactions of the American Fisheries Society; 131: 329-336

Walters, C.J. 1986. Adaptive Management of Renewable Resources. MacMillan, New York.

Walters, C.J. and C.S. Holling. 1990. Large-scale management experiments and learning by doing. Ecology 71: 2060-2068.

Appendix A

This section is intended to explain the rationale for adoption of fatalities/MW/year as the metric of choice as opposed to fatalities/turbine/year. First, the authors criticize the metric of fatalities/turbine/year by stating it can be misleading. Their example on page A-2 demonstrates how fatalities rates using number of turbines as a reference can appear to differ at 2 locations even when the same number of deaths occurs at these locations. The reason for this is the sites have differing numbers of turbines. The same differences can be illustrated using their metric of fatalities/MW/year. For example, if farm A generates 40MW/year and farm B generates only 4MW/year and 100 fatalities occur at both places, then the rate is 2.5 fatalities/MW/year for farm A, but is 25 fatalities/MW/year at farm B. This might mislead someone to believe that more fatalities occurred at farm B. *This issue has been discussed previously.* Later, they compare regression sums of squares relating MW to bird deaths and turbine numbers to support their proposed metric compared to the turbine per year metric. This approach did not compel me to see the advantages of their metric.

The main difference between their metric and the one that uses turbines is that theirs incorporates MW produced per turbine, thus the cost (mortalities) is stated within more of a context of the benefit (MW production). *I have now modified this viewpoint in an earlier description of their defined metric.* Initially, I questioned the purpose of comparing mortality rates among different wind energy generating facilities. If one is only considering management of a particular wind facility, then knowledge of how one place compares to another is not useful. After all, different facilities have many different environmental factors likely to be important in the process that leads to bird strike fatalities. Later, in the discussion, the authors mention replacement of older turbines with larger turbines capable of greater MW generation. Thus, if one wanted to compare fatality rates between time periods at a given site, it may be advantageous to use their metric if considering the cost-benefit aspects of the power generation. They have presented better arguments for using their metric here than previously.

Also in the results section, the authors refer to the relationship between time span surveyed and number of turbines with non-zero mortality (Figure A6). It seems obvious to me that the more you search, the more likely you are to find at least one mortality at a turbine, so I do not understand why they are emphasizing this point as being important. The proportion of turbines where at least one bird has been killed is not a metric that I find particularly useful and does not translate directly to fatalities per turbine or MW per year. I do not agree that this relationship demonstrates that most of their turbines were not sampled long enough to robustly estimate mortality. I am not sure what is meant by 'robustly' here, but in section 4.4.2 on page 86, the authors use the word robust to imply reliability based on high precision. Precision, or repeatability, is only one component of accuracy. Bias, or the deviation of an excepted value of an estimator from the true parameter is equally, or perhaps more, important.

Appendix B

Appendix B attempts to explain the differences in their study from those reported in Kerlinger and Curry (2003). I agree that the incidental counting/voluntary response of bird carcasses is not a rigorous approach to estimating mortality rates if that is what the Wildlife Reporting and Response system relies upon. In various places, the authors refer to 'less robust' estimates of other researchers and then proceed to compare estimates from various studies. Robustness implies a resiliency to, for example, an assumption violation. It is not clear to me what their use of the term implies. In comparing estimates, they report Kerlinger and Curry (2003) underestimated mortality relative to their mortality estimates. Such comparisons are unwise if they constitute different spatial or time periods. Of course, one must know the true mortality to say which estimate is 'better'.

Appendices C and D

These sections consist of details of the chi-square goodness of fit approach to examining use versus availability of various landscape and turbine features. I have previously commented on the efficacy of this approach to analysis.